

CONSIDERING CONSISTENCY
Conceptual and Procedural Guidance
for Reliability in a Local Assessment System

Maine Department of Education

14 June 2004

The Maine Department of Education wishes to thank Theodore Coladarci and Jill Rosenblum for preparing, respectively, Part 1 and Part 2 of this document.



STATE OF MAINE
DEPARTMENT OF EDUCATION
23 STATE HOUSE STATION
AUGUSTA, MAINE
04333-0023

JOHN ELIAS BALDACCI
GOVERNOR

SUSAN A. GENDRON
COMMISSIONER

June 13, 2004

Dear Colleague:

Considering Consistency takes us one step closer to a fully developed system of local assessments, a system built on the conceptual framework established in *Measured Measures* and the procedural guidance of the *LAS Guide*. With *Considering Consistency*, we move toward a more complete understanding of the requirements of scoring assessments, particularly those in the certification set, to ensure reliability of the judgments made regarding student progress toward or attainment of the standards of Maine's *Learning Results*.

Maine has taken a path not traversed by many other states: a system of accountability that is close to the classroom and to teaching and learning. We have adopted this strategy and its accompanying challenges because we believe it will help Maine's teachers and students both enhance learning and document achievement. Throughout the development of *Considering Consistency*, discussions involving the authors and a variety of reviewers have always attempted to balance technical standards with feasibility. This phase of local assessment system development, like all other aspects of the long-term *Learning Results* implementation process, will of necessity need to be evaluated to assess its impacts. The Department has made a major commitment this year to the Local Assessment Implementation Study (LASIS) and I am committed to continuing this approach to "learning as we go."

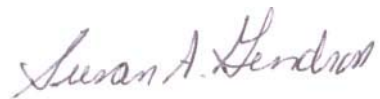
As we move forward to further define concepts and procedures to guide this work, the Department with its advisory committees will explore the merits of developing a model for scorer "certification." This concept highlights teacher judgment supported with a system of checks and balances. Regardless of the nature of developments toward this end, the practices and protocols outlined in *Considering Consistency* will continue to be a necessary foundation for establishing and monitoring reliability in local assessment systems. Undertaking this work will build capacity and document quality under any conceptual model.

It is important to note, as each new local assessment system building block is laid, that both the Department and local educators view local assessment systems as an integral part of—but not a replacement for—other aspects of a comprehensive approach to school improvement. Indeed, the creation of useful assessment data should have as its primary purpose to support effective student, school, and district improvement efforts. To achieve this end, the development of local assessment systems should not be seen as an end in itself, but rather as deeply connected to curriculum revision, instructional program evaluation, data analysis capacity building, student intervention programs, and

public communications, to cite just a few components of a fully developed standards-based system.

I am very grateful that Maine's system of standards and assessments has been built with widespread input and stakeholder involvement. *Considering Consistency* is no exception. A number of individuals have participated in the development and critical review of this document. First and foremost, Ted Coladarci and Jill Rosenblum—contributors to *Measured Measures*—have done their usual outstanding job of framing the issue from both conceptual and operational perspectives. Without their insights into theory and practice, Maine's system would not be as evolved as it is today. In addition, the Technical and Policy Advisory Committees (TAC and PAC) have devoted many hours to evaluating both the details of the guidance contained within these pages and the likely impact in Maine schools once the guidance is implemented. In addition, I wish to thank the educators who participated in Department sponsored focus groups designed to gather critical insights from local practitioners prior to publication. The document was improved based on their comments and suggestions. The students of Maine will be the ultimate beneficiaries of this shared commitment to excellence.

Sincerely,

A handwritten signature in dark ink, reading "Susan A. Gendron". The signature is written in a cursive, flowing style.

Susan A. Gendron
Commissioner

Table of Contents

Introduction.....	4
Part I: Conceptual Guidance	6
Validity vs. Reliability	6
Assessment Reliability vs. System Reliability.....	8
Building Reliability into the System.....	12
Building Reliability into Assessments	13
Assessment Clarity.....	13
Scorer Accuracy	14
Documenting and Monitoring Reliability of Performance Indicator Scores	15
The Relevance of Interrater Agreement in a Local Assessment System	15
The Cumulative Nature of Reliability.....	18
Postscript: Empirically Examining the Consistency of the System.....	20
Part II: Procedural Guidance.....	21
Overview Tool: Steps to Establish the Reliability of a Local Assessment System	23
Suggested Models for Scoring and Establishing Reliability.....	24
Preparing for Training, Calibration, & Scorer “Certification”	31
Tool Set #1—Appropriate Selection and Distribution of Assessments for an LAS.....	32
Template for English Language Arts.....	33
Template for Health and Physical Education.....	34
Template for Mathematics	35
Template for Science & Technology	36
Template for Social Studies	37
Template for Career Preparation.....	38
Template for Modern and Classical Languages.....	39
Template for Visual and Performing Arts	40
Sample Template for Science & Technology	41
Tool Set #2—Checking the Quality of Assessments	42
Clarity of Language	43
Anatomy of a Scoring Guide	44
Guidelines for Drafting Scoring Guides	45
Steps for Drafting Scoring Guides.....	46
Scoring Guide Template (multiple performance indicators)	47
Scoring Guide Template (single performance indicator).....	48
Sample Scoring Guide (“Patterns, Relations, Functions”)	49
Sample Scorer’s Notes (“Patterns, Relations, Functions”).....	50
Tool Set #3—Developing and Ensuring Scorer Accuracy	51
Ground Rules for Scoring Student Work.....	54
Scorer Bias Considerations	56
“Not Scorable” Guidelines.....	59
Quick Tips for Scoring Student Work	61
Guidelines for Selecting Benchmarks.....	62

Guidelines for Writing Commentary or Scoring Rationales for Benchmarks or Other Training Materials.....	63
Template for Scoring Rationale/Commentary	64
Sample Scoring Rationale/Commentary	65
Facilitating Training and Calibration Sessions	66
Guidelines for Selecting Student Work for Double Scoring	69
Appropriate Sample Size & Interrater Agreement for Double Scoring.....	70
Double Scoring Student Work	71
Guidelines for Establishing Reliability for Assessments involving Observation, Presentation, or Performance.....	72
Sample Interrater Agreement Tally	73
Recommendations for Addressing Insufficient Interrater Agreement.....	75

Introduction

The Maine State Legislature established the *Learning Results* in 1996, thereby ushering in the age of standards-based education in Maine. The legislature also stipulated that student achievement of the standards was to be measured by a combination of state and local assessments, which gave rise in Maine to the notion of a “local assessment system.” Chapter 127, a Department of Education Regulation amended in 2002, delineates standards for local assessment systems and individual assessments alike. This was followed by *LAS Guide: Principles and Criteria for the Adoption of Local Assessment Systems* (Maine Department of Education, 2003), which established the required framework for developing a local assessment system.

After the release of the *LAS Guide*, many Maine educators sought guidance regarding the reliability requirements for a local assessment system. Wasn’t this supposed to be addressed in *Measured Measures*,¹ you reasonably may ask? Released in 2000, *Measured Measures* provided technical guidance for developing local assessment systems. Admittedly, not much was known then about the technical implications of a *system* of assessments. Consequently, *Measured Measures* focused more on the individual assessment. For example, we described ways to estimate the reliability of an assessment, but we remained silent on how to think about reliability within the context of a system. Similarly, we focused on the validity of an assessment, but we made only passing reference to the validity of a system. These are not subtle distinctions: It is the difference between establishing confidence in our judgment about a student based on a single assessment versus establishing confidence in our ultimate certification decision based on the

¹ Coladarci, T., Johnson, J. L., Beaudry, J., Cormier, M., Ervin, R., Rosenblum, J. M., & Silvernail, D. L. (2000). *Measured Measures: Technical Considerations for Developing a Local Assessment System*. Augusta, ME: Maine Department of Education. [<http://www.state.me.us/education/g2000/measured.pdf>]

collection of assessments across the local assessment system. By establishing confidence in the latter, one is able to speak to the defensibility of these important judgments.

Our purpose in writing *Considering Consistency* is to provide conceptual and procedural guidance regarding reliability within the context of a local assessment system. Some of what we offer will echo *Measured Measures*. For instance, we revisit methods for estimating the reliability of individual assessments, particularly as appropriate for constructed-response and performance-based assessments. But unlike *Measured Measures*, the present document offers specific guidance regarding desirable levels of reliability and, moreover, it does so by considering the system as a whole. For example, what is meant by the “reliability” of a local assessment system? What strategies are available for establishing reliability in this regard? What is an acceptable reliability at the lowest level in the system? How does reliability change as one moves from one level in the system to the next—e.g., from the performance indicator score to the content cluster to the content area?

We have separated *Considering Consistency* into conceptual guidance (Part I) and procedural guidance (Part II). Our intention, of course, is that the conceptual discussion provides a helpful rationale for what follows procedurally. But we also know that by better understanding the procedural, readers in turn will have a fuller appreciation of the conceptual. (Indeed, some readers may prefer to approach the document in reverse order.) We believe any educator can benefit from a careful reading of *Considering Consistency*. Nevertheless, we envision that, locally, the task of becoming thoroughly acquainted with *Considering Consistency* will fall primarily on the shoulders of a small team who, in turn, will use the document as a framework for staff development, local assessment initiatives, and related activities.

PART I: CONCEPTUAL GUIDANCE

We begin by comparing the concepts of validity and reliability, after which we examine the meaning of reliability at different levels in a local assessment system. We then address how one builds reliability into the system and, in turn, how reliability is built into individual assessments. After this, we discuss reliability as applied to performance indicator scores, which are the building blocks of a local assessment system as established by the *LAS Guide*. Finally, we offer a postscript regarding system-level analyses down the road. Throughout, you will find reference to “tool sets” that will follow in Part II.

Validity vs. Reliability

Measured Measures treated validity and reliability separately. We began with validity, which we defined as “the extent to which an assessment measures what it is supposed to measure, and the extent to which inferences and actions on the basis of tests scores are appropriate and accurate” (p. 7).² For an individual assessment, the first part of this definition means that the assessment demonstratively aligns with the targeted *Learning Results* performance indicators—that the assessment in fact calls for the knowledge, skills, or processes represented by the targeted indicators.

It is equally important, particularly with constructed response items, that the assessment also is scored for the knowledge, skills, or processes represented by these indicators. If a scorer is using a 4-point rubric and assigns, say, a 3 (e.g., “meets the standard”) to a piece of student work, this score as applied to this piece should unassailably correspond to the defining characteristics of student work, that in, fact meets the standard. Similarly, each of the other

² We take this definition from the online assessment glossary of the National Center for Research on Evaluation, Standards, and Student Testing (<http://www.cse.ucla.edu/CRESST/pages/glossary.htm>).

assigned scores should reflect student work that is consistent with the respective performance level descriptions. In short, the ability of scorers to “score to standard”—that is to say, to score accurately—is an important aspect of validity.

The second part of the validity definition—“the extent to which inferences and actions on the basis of tests scores are appropriate and accurate”—means that assessment-based inferences (e.g., conclusions about a student’s knowledge or ability) and assessment-based actions (e.g., decisions regarding instruction or placement) are in keeping with these performance indicators. For example, if your assessment is designed to measure mathematical computation skills, then you would limit your inferences and actions to those relevant to this domain.

We then turned to reliability, or “the consistency of the scores, ratings, or judgments that derive from an assessment” (*Measured Measures*, p. 19). For example, you would question the reliability of your weight scale if you obtained discrepant results upon weighing yourself twice (identically clothed) within a 10-second interval, just as you would question the reliability of proficiency judgments if two teachers, having examined the same pieces of work, gave widely different ratings for many students.

It is important to understand that “reliability” does not entail “validity.” An assessment high in reliability (e.g., a physician’s weight scale) nonetheless can be low in validity for a particular purpose (e.g., making inferences about verbal ability). Nevertheless, and despite the separate treatment of reliability and validity in *Measured Measures*, the distinction between the two concepts in fact can be quite subtle. Indeed, there is the growing recognition that reliability is an aspect of validity rather than a separate consideration altogether.³ For instance, if an assessment does not yield consistent (reliable) results, then these results necessarily limit the

³ Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). New York: Macmillan.

corresponding inferences that can be made about student learning (validity). Or consider the notion of scoring to standard, which we introduced above. If scorers demonstrate the ability to score to standard (validity), in doing so they are demonstrating their ability to score consistently (reliability). Scorers who score accurately are necessarily scoring consistently.

The close relationship between reliability and validity also is evident with respect to a local assessment system as established in the *LAS Guide*. By building a system of assessments that exemplifies the criteria of *coherence* and *sufficiency*, the school administrative unit thereby shores up the reliability of the system. But a coherent and sufficient system simultaneously permits clearer and more defensible inferences about student learning as well. In short, coherence and sufficiency have as much to do with validity as they do reliability. (As you will see, Tool Set #1 pertains to coherence and sufficiency.)

Assessment Reliability vs. System Reliability

Whether in regard to a single performance indicator score or the system as a whole, reliability fundamentally reflects consistency. But by raising the distinction between *assessment* reliability and *system* reliability, we underscore the hierarchical nature of information in a local assessment system (see Figure 1). As established in the *LAS Guide*, the most elemental level in this hierarchy is a measurable outcome: a single 4-point score tied to a *Learning Results* performance indicator. We refer to this as a performance indicator score, and it takes on values ranging from 1 to 4. The next level up is a collection of performance indicator scores across a cluster of content standards in the content area. And, ultimately, we have the aggregation of scores across the entire content area.

You may wonder why we have neglected to mention the *assessment* in this hierarchy. Because an assessment may span multiple clusters in a local assessment system, just as a cluster may be relevant to only part of an assessment, “the assessment” doesn’t fit neatly into the hierarchy shown in Figure 1. As established by the *LAS Guide*, performance indicator scores are the building blocks of the local assessment system. They are aggregated up to the content cluster and, in turn, to the content area as a whole.

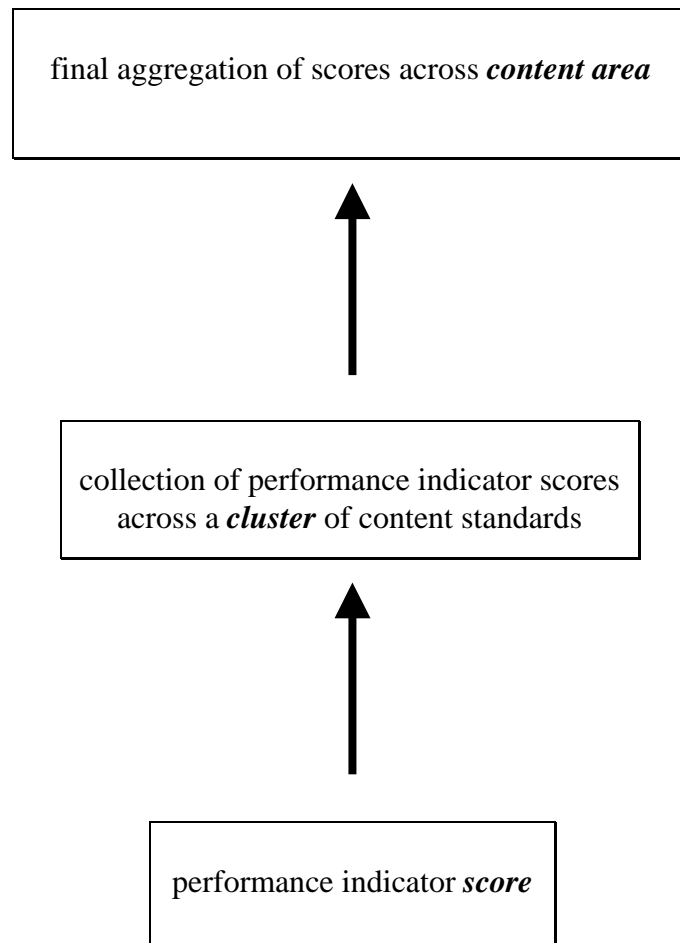


Figure 1. The hierarchical nature of a local assessment system.

One can pose the reliability question at each level of Figure 1—inquiring about the reliability of a single performance indicator score all the way to the reliability of the final certification judgment in the content area. The importance of reliability increases with each level in the system, given the corresponding hierarchy of consequences. As shown in Figure 2, performance indicators scores are most appropriate for making low stakes, easily modified, classroom instructional decisions. Consequently, reliability here can be comparatively modest. Cluster level data are most helpful for monitoring programs and adjusting curriculum, in which case a higher reliability is desired. Finally, a still higher reliability is required across the entire content area because of the high-stakes certification decisions made from data at this level of the system.

Establishing reliability at the most elemental level—the performance indicator score—is rather straightforward, as we illustrate later. But demonstrating reliability at this level does not establish the reliability of the *system* of local assessments, any more than having five gifted musicians will ensure a melodic composition. In the latter case, the five virtuosos must *cohere* as a quintet. There similarly must be coherence among the various assessments (and the performance indicator scores they comprise) in a local assessment system.

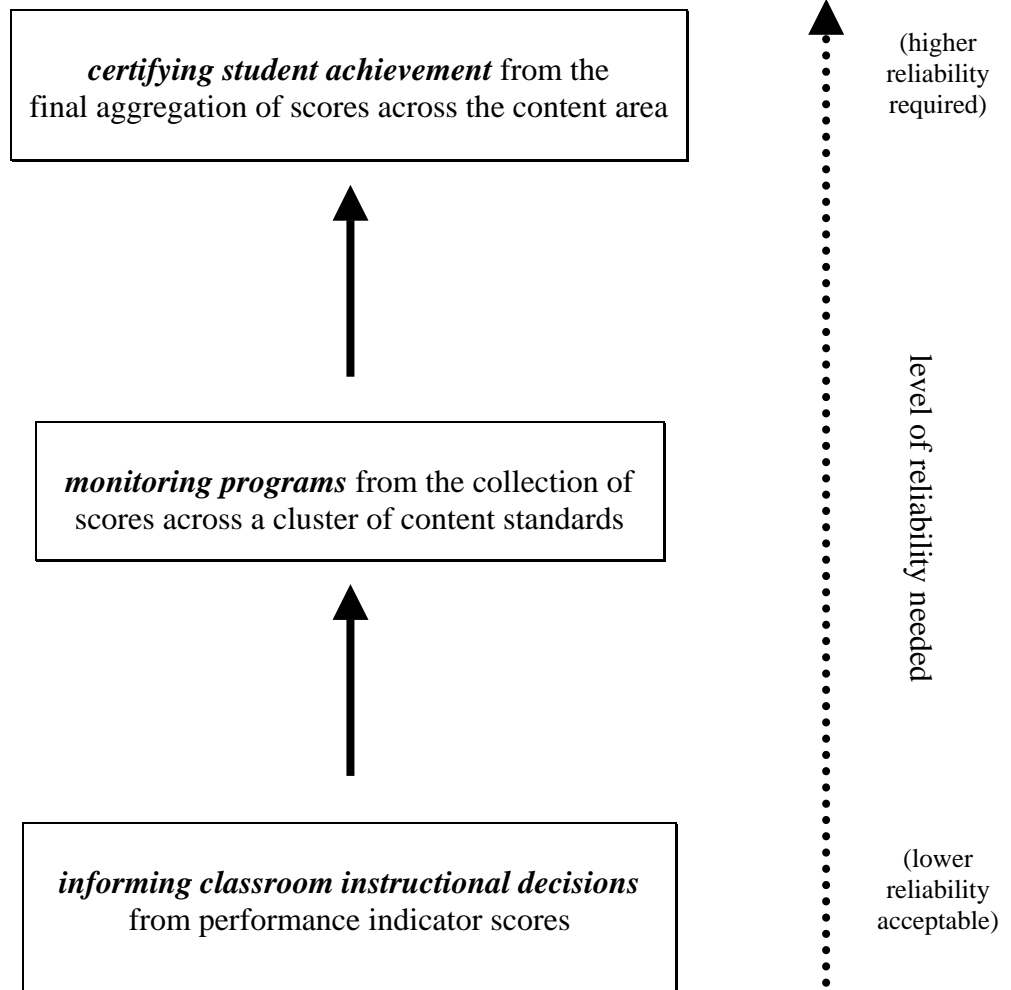


Figure 2. Reliability and purpose in a local assessment system.

Building Reliability into the System

Unfortunately, there is no straightforward and practical way to statistically demonstrate the coherence of a local assessment system. Rather, the coherence of a local assessment system is established largely *by design*. As discussed in the *LAS Guide* (p. 5), this is accomplished by designing or selecting assessments that, collectively, “function in an integrated and balanced fashion within the system” and are “representative of both the discipline and of the students’ skills and knowledge.” Toward this end, the deliberative selection of assessment types and their strategic distribution across the content area are both of central importance (*LAS Guide*, pp. 9-13).

The “size” of the system also is important to its reliability. Just as test reliability generally increases with test length, the reliability of a local assessment system is related to the amount of information the system comprises. It is for this reason that the *LAS Guide* calls for at least 8 assessments in a grade span per content area and, moreover, a minimum of 5 performance indicator scores for each cluster in a content area. These minima are intended to build *sufficiency* into a local assessment system.

Tool Set #1, which you will find in Part II of this document, provides procedural guidance for building coherence and sufficiency into the assessment system through the thoughtful selection and distribution of assessments and assessment types.

Finally, system reliability is bolstered by the consistent application of a common set of performance standards when making the final certification decision. Further, just as coherence and sufficiency are related to reliability and validity alike, so too are performance standards:

They contribute to system validity by being rooted in standards-based declarations of what is “good enough,” given student performance on assessments that align with the *Learning Results*.

By designing a local assessment system that is characterized by coherence, sufficiency, and the consistent application of performance standards, the school administrative unit is building reliability (as well as validity) into its system. Once a local assessment system is well established and has generated at least a grade-span’s worth of data, one then can conduct various analyses to empirically explore the reliability (and validity) of the system. We conclude Part 1 with a sketch of some possible analyses toward this end.

Building Reliability into Assessments

Just as the school administrative unit builds reliability into its system by design, each assessment is crafted or selected to enhance reliability. Assessment clarity and scorer accuracy are two factors that influence the reliability of an assessment over which the school administrative unit has direct control.

Assessment Clarity

- Reliability is enhanced when the assessment has clear and unambiguous language: clear constructed-response questions, clear writing prompts, clear instructions for performance assessments and exhibitions, clear selected-response items, and so on. Ambiguity introduces a degree of randomness in student performance, which lowers reliability.

- Reliability is enhanced by clarity of scoring criteria: When criteria for evaluating student performance are stated in a clear and straightforward manner. For constructed-response and performance-based assessments, this calls for a well-specified scoring guide. This is of particular relevance for a local assessment system, where the most elemental unit of assessment is the 4-point performance indicator score.

Tool Set #2, in Part II of this document, provides procedural guidance for evaluating the clarity of assessments and, further, for drafting standards-based scoring guides designed to support consistent judgments about student performance.

Scorer Accuracy

It is not enough to have a good scoring guide. Scorers must apply these criteria consistently and with fidelity. This calls for scorer training and calibration, initial reliability checks, subsequent monitoring of reliability, and occasional recalibration. The more emphasis that is placed on initial training and calibration, the greater the impact on scorers' ability to score to standard and, in doing so, to score consistently. During actual scoring sessions, score behinds and double scoring are essential for monitoring scorer accuracy and consistency.

Tool Set #3, in Part II of this document, provides procedural guidance for planning and running scoring sessions: sample materials, guidelines for selecting a sample for double scoring, strategies for training and calibrating scorers, and tools for documenting interrater agreement.

Documenting and Monitoring Reliability of Performance Indicator Scores

Because performance indicator scores are the building blocks of the local assessment system, we focus on the performance indicator score in offering guidance for documenting and monitoring reliability. Why not the system as a whole, you may ask? As we argued earlier, system reliability is best approached *by design*.

In short, our contention is this: If each assessment is designed or selected to contribute both coherence and sufficiency to the local assessment system, if performance indicator scores are of demonstrable reliability, and if a common set of performance standards are consistently applied, then a school administrative unit can be confident in the reliability of its local assessment system.

The Relevance of Interrater Agreement in a Local Assessment System

As you know by now, reliability reflects consistency. But “consistency” can take various forms. Depending on the assessment context, it can mean:

- consistency across *raters*
- consistency across *items*
- consistency across *forms*
- consistency across *occasions*

Table 1 briefly elaborates on each aspect of consistency and its relevance to a local assessment system. The first aspect, consistency across raters, arguably is the most relevant in the present context, given the centrality of 4-point scores in a local assessment system. Provided that these scores are determined within the context of a coherent and sufficient system, the fundamental reliability question is one of interrater agreement. That is, do independent raters of

the same body of student work, using the same scoring guide, arrive at similar judgments? By demonstrating that there is adequate agreement in this regard, the school administrative unit is speaking to the reliability of these scores. Again, the reliability of the system is established when demonstrably reliable performance indicator scores are coupled with a local assessment system that is characterized by coherence, sufficiency, and the consistent application of performance standards.

Although interrater agreement is our focus in *Considering Consistency*, one should not infer that interrater agreement is the only appropriate form of reliability in the present context. As Table 1 indicates, internal-consistency reliability would be appropriate for an assessment having selected-response items, and equivalent-forms reliability would be appropriate where two versions of an assessment are available (e.g., for replacement purposes, or rotation for security purposes). But again, our emphasis here is on interrater agreement because of the centrality of 4-point scores in a local assessment system.

Table 1. Reliability: The different facets of “consistency”

Consistency across raters

- Context: e.g., a constructed-response assessment, scored with a 4-point rubric.
- Reliability question: Do two raters, using the same rubric and independently, judging a common sample of student work, classify each piece of work the same?
- Reliability index: interrater agreement.
- Relevance to LAS: High (applicable to any performance indicator score).

Consistency across items

- Context: e.g., a 50-item selected-response assessment of content knowledge in a particular discipline.
- Reliability question: Is the assessment test internally consistent? In other words, is there evidence that each item contributes meaningfully to the total score?
- Reliability index: internal consistency (e.g., Cronbach’s alpha, Kuder-Richardson).
- Relevance to LAS: Applicable to assessments having selected-response items.⁴

Consistency across equivalent forms

- Context: e.g., a school administrative unit has two equivalent, or parallel, versions of an assessment (e.g., to allow for replacement or rotation for security); student performance is evaluated using a 4-point rubric.
- Reliability question: If the same group of students were to take both versions of this assessment, would the two assessments yield similar classifications of students?
- Reliability index: classification consistency⁵.
- Relevance to LAS: Applicable where there are two versions of an assessment.

Consistency across occasions

- Context: e.g., an assessment purportedly measures knowledge, skills, etc. regarded to be “stable” over time.
- Reliability question: If the assessment were given to the same group of students on two occasions, would the rank order of student performance be similar (stable) across the two occasions?
- Reliability index: correlation coefficient⁶.
- Relevance to LAS: Low, insofar as “stability” in performance is not central to standards-based education.

⁴ See *Measured Measures* for the assumptions required by this form of reliability (p. 22) and for a worked example (Appendix E).

⁵ See Appendix D in *Measured Measures* for a worked example of classification consistency involving a dichotomous judgment (proficient vs. not proficient).

⁶ Classification consistency would be appropriate here if, in the reliability question, you replaced *rank order* with *classification*.

We recommend exact agreement of 70% or greater for each 4-point performance indicator score. Although any recommended level of reliability necessarily has a degree of arbitrariness, 70% is consistent with prevailing views of acceptable interrater agreement.⁷ More importantly, in the present context this value strikes us as neither too low nor too high. On the one hand, 70% exact agreement is high enough for informing day-to-day classroom instructional decisions regarding a student or group, where consequences are relatively minor and decisions can easily and quickly be modified.⁸ On the other hand, this value is not so high that it places an unnecessary, and possibly unrealistic, expectation on the school administrative unit.

As we describe in Part II, interrater agreement is estimated by double scoring a representative sample of all papers resulting from a particular assessment. Because only a sample of papers is double scored (rather than all of them), it is important that the sample is large enough to provide confidence that the calculated interrater agreement is indicative of the degree of scoring consistency among all papers. Specific guidelines are offered in Part II, which attempt to balance concerns regarding technical sufficiency with those regarding feasibility.

The Cumulative Nature of Reliability

Perhaps you are thinking that the recommended value—70% exact agreement—is a bit modest for assessment results that ultimately will be used for making certification decisions. Bear in mind that the recommended value is for a single performance indicator score. When individual performance indicator scores are aggregated up to the cluster, reliability will be

⁷ Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, 9(4). [Retrieved March 1, 2004 from <http://PAREonline.net/getvn.asp?v=9&n=4>]

⁸ For example, see Nitko, A. J. (2001). *Educational assessment of students* (3rd ed.; pp. 76-77). Upper Saddle River, NJ: Prentice-Hall.

considerably higher. And reliability across the entire content area—where the certification decision is made—will be higher still.

Why is this so? As we intimated earlier, an axiom in educational assessment is that test reliability generally increases with test length. This is because a longer assessment (e.g., many items) provides students a better opportunity to demonstrate what they know and can do than is possible on a shorter assessment (e.g., a single item). As a consequence, student performance on the longer assessment offers a more reliable—a more dependable—indicator of their standing on whatever is being assessed. It is like the greater confidence you would have in a claim about popular opinion on some issue when the claim is based on a random sample of 1,500 people rather than only 5, or the greater confidence you would have in a judgment of a tennis player's serve after an entire match rather than after the first toss. Just like a longer test, a larger random sample, or an entire tennis match, a higher level in the local assessment system (where there is more information) will have superior reliability to that at a lower level in the system (where there is less information).

Thus, reliability is cumulative in a local assessment system that has coherence and sufficiency. Although you would be only moderately confident in a judgment about a student at the performance indicator level, you would have considerably more confidence in a cluster-level characterization of this student, and you would have even more confidence in the ultimate judgment regarding certification in the content area for this student.

Tool Set #3, in Part II of this document, includes procedural guidance for when interrater agreement on a performance indicator rating falls below 70%.

Postscript: Empirically Examining the Consistency of the System

As we suggested above, the reliability (and validity) of a local assessment system can be studied more comprehensively once the system has generated at least a grade-span's worth of data. For example, one could examine the intercorrelations among (or generalization across) the various assessments in the system. Where an external measure, such as the Maine Education Assessment (MEA), is included in a local assessment system, one also could examine classification consistency. For instance, performance-level classifications based solely on local assessments could be compared with those based solely on the MEA: Are students classified similarly? Clearly, school administrative units will benefit from additional guidance regarding such analyses.

PART II: PROCEDURAL GUIDANCE

We now turn to the procedural, where we provide tool sets comprising guidelines and templates for developing and documenting reliability in a local assessment system (LAS). There are three tool sets:

- Tool Set #1: Appropriate Selection and Distribution of Assessments for an LAS
- Tool Set #2: Checking the Quality of Assessments
- Tool Set #3: Developing and Ensuring Scorer Accuracy

Each tool set is presented in a separate section. A tool set section includes an explanation of the purpose and uses of the guidelines and templates in the set, a list of the tools, suggestions for their implementation, camera-ready copies and, where appropriate, completed samples or annotated versions of these tools.

Before these tools sets, we begin this section by providing an overview to guide the step-by-step procedures necessary to establish reliability. This is followed by six scenarios describing various models for organizing and conducting the scoring of assessments and by set of recommendations for preparing training materials.

Tool Set #1 contains materials tool for documenting the depth and breadth of the collection of assessments selected for inclusion in the local assessment system. Then we offer guidelines for reviewing assessment quality and templates for drafting scoring guides that support reliability (Tool Set #2). We next address how one organizes and runs a calibration and scoring session, including examples of materials to be used (Tool Set #3). This includes guidelines for selecting a sufficient and appropriate sample of papers for double scoring. These are followed by a tool to use in calculating interrater agreement and to reflect on issues that may be interfering with achieving necessary levels of reliability. The final tool provides a list of

follow-up steps to be implemented when reliability, as estimated by interrater agreement, is not achieved.

Many of these tools, templates, and guidelines are versions of others used in workshops and informational sessions offered by the Department of Education over the last few years. These have been developed and refined over time by many members of the Department of Education and Maine Mathematics and Science Alliance staffs, and with contributions by Maine educators who have participated in our assessment efforts. Others are newly drafted for *Considering Consistency*. We have included all of the relevant tools here—old, revised, and new—so that they all will be easily available as school districts proceed with the work of ensuring the reliability of their local assessment systems. We hope that the convenience of the collection, as well as the explanations and directions, will support both the understanding and accomplishment of reliability.

OVERVIEW TOOL

Steps to Establish the Reliability of a Local Assessment System

This list outlines the reliability related activities necessary in the development and implementation of a Local Assessment System. Where appropriate, we refer to specific tools found elsewhere in this document. On the pages following this overview, we provide descriptions of a number of scoring models designed to accomplish the necessary work of scoring and establishing the reliability of scores.

1. Make initial assessment choices. Select, adapt, and/or develop assessments to meet the requirements of coherence and sufficiency (see Tool Set #1 pages 33 – 40).
2. For assessments that have undergone a formal field test and have demonstrated reliability, proceed to Step 4.
3. Assessments that have been locally developed or adapted must be field tested and scored to finalize the scoring guide (see Tool Set #2 pages 45 and 96), select and annotate benchmarks/anchor papers (see Tool Set #3 pages 62 – 64), and document acceptable levels of interrater agreement (see Tool Set #3 pages 72 – 73).
4. Implement assessments as scheduled within LAS.
5. Plan and hold scorer training including review of scoring guide and benchmarks/anchor papers, practice scoring, and calibration activities (see Tool Set #3 page 66).
6. Complete scoring of all student work including a strategy to establish the reliability of the scoring (through interrater agreement), or the scorer (through scorer “certification”). See suggested models on the following pages 24 – 30.
7. When adequate interrater agreement is not achieved, additional training and rescoring will be necessary (see Tool Set #3, pages 70 and 74).

Suggested Models for Scoring and Establishing Reliability

Select one of the models below, or design another strategy or variation that provides evidence of reliability through interrater agreement or documentation of scorer “certification”.

- a. All teachers meet together to score student work and distribute sample for double scoring throughout the session.

For example: All of a district’s third grade teachers meet on a staff development or release day for training and calibration, and to score an assessment. Teachers avoid scoring their own students’ work and 25 pieces of student work are double scored to document levels of interrater agreement. Adequate interrater agreement documents the reliability of scoring completed during the session. Throughout the session, a facilitator “scores behind” each scorer to check for consistency and accuracy. When disagreement is detected, the facilitator provides individual training to recalibrate the scorer.

Advantages

- Enhances professional development opportunity and supports collegial discussions about implications for curriculum and instruction.
- Allows for “score behinds” and ongoing retraining, especially important early in the process as educators develop scoring skill and capacity.
- Avoids teachers scoring their own students’ work, eliminating a potential source of bias.

Considerations

- Requires large block(s) of release or inservice time.
- Works best with group of several teachers. This group may include others in a grade span, or other educators in a school—beyond those teachers who actually administered the assessment.

- b. Teachers score their own students' work and bring identified (randomly selected) sample to a second scoring session.

For example: all of a district's fifth grade teachers (six teachers) score their own students' work sometime between the scorer training and calibration (offered during common planning time, during an early release, or after school) and the second scoring session (also could be scheduled during common planning, early release, etc.). Each of the teachers brings five randomly selected student papers- for example the 1st, 7th, 10th, 16th, and 21st. These are distributed, one each, to the five colleagues. Each teacher scores five papers, one from each of the other fifth grade teachers. Adequate interrater agreement documents the reliability of the scoring that teachers have done.

Advantages

- Provides flexibility in finding scoring time.
- Does not require large block(s) of release or inservice time.

Considerations

- Relies on initial training and calibration to ensure consistent application of scoring guide (no "score behinds").
- Limits opportunities for collegial discussions.
- Requires teachers to score their own students work and to avoid being biased by their knowledge and feelings about the students.

- c. Teachers meet in a regional group to score student work and distribute a sample for double scoring throughout the session.

For example: Three physics teachers from three different high schools agree to administer the same assessment to their students. While substitutes cover their classes, the three meet for training and calibration and to score the work from all of their students. The teachers avoid scoring their own students' work and 25 pieces of student work are double scored to document levels of interrater agreement. Adequate interrater agreement documents the reliability of scoring completed during the session.

Advantages

- Addresses situations in small schools or departments where there is only one teacher at a particular grade level or for a particular subject area.
- Expands professional development opportunity and supports regional collegial discussions about implications for curriculum and instruction.
- Allows for “score behinds” and ongoing retraining.
- Avoids teachers scoring their own students' work, eliminating a potential source of bias.

Considerations

- Involves logistics and planning for regional meeting.
- Requires large block(s) of release or inservice time.

- d. Teachers score their own students' work and bring identified (randomly selected) sample to regional second scoring session.

For example: Three physical education teachers from different schools agree to administer the same assessment to their students. They meet after school one day to prepare for scoring their own students' work by reviewing the scoring guide and the videotaped benchmark samples and by scoring some additional videotapes (training and calibration). The teachers score their students as they administer the assessment and videotape a sample to bring for second scoring. The three teachers meet again to view and score videotapes from one another's schools. Adequate interrater agreement documents the reliability of the scoring that teachers have done.

Advantages

- Addresses situations in small schools or departments where there is only one teacher at a particular grade level or for a particular subject area.
- Allows for some regional collegial discussion.
- Provides flexibility in finding scoring time.
- Requires shorter block of meeting time.

Considerations

- Involves logistics and planning for regional meeting.
- Relies on initial training and calibration to ensure consistent application of scoring guide (no "score behinds").
- Requires teachers to score their own students work and to avoid being biased by their knowledge and feelings about the students.

- e. Teachers score their own students work and provide a SAU scoring team with samples to be double scored.

For example: All of the SAU's sixth grade teachers administered a writing prompt. They meet to review the scoring guide and complete training and calibration activities. Then each teacher scores his or her own students' writing pieces. Each teacher submits several papers to the SAU Writing Assessment Committee. The committee meets, reviews the scoring guide and benchmarks, completes calibration activities and then scores the 25 papers submitted by the sixth grade teachers. Adequate interrater agreement documents the reliability of the scoring that teachers have done.

Advantages

- Provides flexibility in finding scoring time.
- Does not require large block(s) of release or inservice time.
- Builds leadership and assessment capacity of committee members.

Considerations

- Relies on initial training and calibration to ensure consistent application of scoring guide (no “score behinds”).
- Limits opportunities for collegial discussions.
- Requires teachers to score their own students work and to avoid being biased by their knowledge and feelings about the students.
- Depends on the existence and support (time and compensation) of a committee with the skills and knowledge necessary to provide second scores on a variety of assessments.

- f. One or more teachers documents scorer “certification” for an assessment and then scores his or her own students’ work on that assessment.

For example: All of an SAU’s first grade teachers agree to administer a common science assessment on life cycles but some will use it as the culminating activity for a fall unit on monarch butterflies and others to wrap up a spring unit on tadpoles and frogs. Since the assessment will be administered at different times, as each teacher gives the assessment, he or she makes time to prepare for scoring. The teacher logs on to the Maine Assessment Portfolio (MAP) website, reviews the scoring guide and benchmarks and completes the practice scoring pieces. Then the teacher scores the set of 10 calibration pieces provided on the site. The website checks the teacher’s scores against the “true scores” assigned by a statewide group of teacher leaders. If the teacher has 80% or better agreement, then he or she has been “certified” for that assessment. After documenting scorer “certification”, each teacher scores his or her own students’ work.

Advantages

- Verifies accuracy, as well as consistency, of scorer.
- Addresses situations in small schools or departments where there is only one teacher at a particular grade level or for a particular subject area.
- Attends to situations where a common assessment is administered at different times to different students or groups of students.
- Allows for individual teachers to score performances or presentations without requiring videotaping for second scorer.
- Provides flexibility in finding training, calibration, and scoring, time.
- Does not require large block(s) of release or inservice time.

Considerations

- Limited to those MAP assessments that provide online, interactive scoring practice and calibration OR,
- Requires substantial investment of time to prepare training and calibration sets with “true” scores and commentary to be used for documentation of individual scorer “certification”.
- Relies on initial training and calibration to ensure consistent application of scoring guide (no “score behinds”).
- Eliminates opportunities for collegial discussions.
- Requires teachers to score their own students work and to avoid being biased by their knowledge and feelings about the students.
- Requires strategy for addressing issues arising when a teacher fails to become “certified”.

General Recommendation

We believe that a combination of the strategies listed above will represent the best way to address the need to score and document reliability on all of the assessments in an LAS. This allows SAUs maximum flexibility in creating efficient, workable systems and making modifications based on local needs and available resources.

Choices about which strategy to use (and when) should take into account:

- **teacher experience and capacity**
- **availability of release time, inservice time, common planning time, etc.,**
- **teacher leadership capacity (facilitation, training, “score behinds”)**
- **assessment selection, including availability of calibration/certification materials.**

Strategies are likely to change over time as capacity increases, the LAS is modified, and schedules shift.

Preparing for Training, Calibration, & Scorer “Certification” Suggestions for Building Capacity and Compiling Materials

This list includes ideas for developing teacher leaders who are capable of facilitating training and scoring sessions and for collecting and preparing the materials necessary for scorer training and calibration, and, potentially, for scorer “certification”.

DEVELOPING TEACHER LEADERSHIP

- Identify teachers with an interest in assessment and/or commitment to LAS development and implementation
- Consider teachers with experience through MAP and/or LAD participation
- Engage teacher leaders in planning training and calibration sessions – create template
- Involve teacher leaders in preparing benchmarks, training, and calibration materials
- Establish accuracy and consistency of teacher leaders’ scoring – “certification”
- Provide training in facilitation, read behinds, moderation/retraining

COLLECTING MATERIALS FOR TRAINING, CALIBRATION, AND “CERTIFICATION”

In addition to benchmark samples with commentary, each assessment that is part of the LAS should be accompanied by a collection of accurately scored student work for use in training and calibration activities, and scorer “certification”. These collections may have to be developed over time.

- Select up to 20 samples of student work for each assessment
- Include samples at each of the four score points, and not scorable (NS) – in proportion to their actual occurrence
- Include samples that represent “borderlines” or “tough calls”
- When there is a choice, select samples that will reproduce well
- When there is no choice, trace student work to ensure that it will reproduce clearly
- Verify all scores on each piece
- Write commentary explaining and justifying each score on each piece and providing specific examples of evidence (see Guidelines for Writing Scoring Commentary, page 63)

TOOL SET #1—Appropriate Selection and Distribution of Assessments for an LAS

This is the first step in building reliability into the LAS. As described in Part I, adhering to the *LAS Guide* requirements for selection and distribution of assessment types contributes to the coherence and sufficiency of the system, which, in turn, builds reliability into the LAS.

Each template in this tool set includes the content standards for a particular subject area, divided into content clusters, at a particular grade span. By completing each template, one checks the collection of assessments for overall number (at least 8), distribution across the content standards (at least one addressing each) and within the content clusters (at least five addressing each), and variety in assessment type within content clusters.

There are templates for English Language Arts, Health and Physical Education, Mathematics, Science and Technology, Social Studies, Career Preparation, Modern and Classical Languages, and Visual and Performing Arts. Each template provides options for use at the K-4, 5-8, or 9-12 grade span; a separate template should be used for each.

Complete the templates by listing the assessment title, the source of the assessment (including, as appropriate, “locally developed,” “adapted from LAD,” “MAP,” etc.). Note the assessment type for each and indicate the content standards that the assessment addresses. If an assessment is scored for more than one performance indicator within a content standard, double check that cell in order to accurately document the number of pieces of evidence available for each content standard. After completing a collection, check the columns to ensure that

- there are at least 8 assessments (per grade span per content area),
- each content standard is addressed at least once (at least 1 performance indicator rating for each),
- each content cluster is addressed at least 5 times (at least 5 performance indicator ratings for each),
- and there is a variety of assessment types for each cluster included in the collection.

**Tool Set #1 Appropriate Selection and Distribution of Assessments for an LAS
Template for English Language Arts**

Grade Span (check one) ___ PK-4 elementary ___ 5-8 middle level ___ 9-12 secondary				Reading and Viewing Cluster			Writing and Speaking Cluster			Integrated Literacy Cluster	
Assessment Title	When Given	Source of Assessment	Assessment Type	A	B	D	E	F	G	C	H
Total # of Assessments Minimum 8-12			Total # of Assessment Types Variety of types/cluster	# Measures ▪ Min 1/ Standard ▪ Min 5/ Cluster ▪ Variety of Types			# Measures ▪ Min 1/ Standard ▪ Min 5/ Cluster ▪ Variety of Types			# Measures ▪ Min 1/ Standard ▪ Min 5/ Cluster ▪ Variety of Types	

**Tool Set #1 Appropriate Selection and Distribution of Assessments for an LAS
Template for Health and Physical Education**

Grade Span (check one) <input type="checkbox"/> PK-4 elementary <input type="checkbox"/> 5-8 middle level <input type="checkbox"/> 9-12 secondary				Health Education Content (Topic) Area Groupings Minimum 1 measure per content (topic) area grouping	Health Knowledge			Health Skills			Physical Education Knowledge and Skills			
Assessment Title	When Given	Source of Assessment	Assessment Type		A	B	D	C	E	F	A	B	C	
Total # of Assessments Minimum 8-12				Total # of Assessment Types Variety of types/cluster	Total # of content (topic) areas groupings Minimum 1 measure per content (topic) area grouping	# Measures <input type="checkbox"/> Min 1/ Standard <input type="checkbox"/> Min 5/ Cluster <input type="checkbox"/> Variety of Types			# Measures <input type="checkbox"/> Min 1/ Standard <input type="checkbox"/> Min 5/ Cluster <input type="checkbox"/> Variety of Types			# Measures <input type="checkbox"/> Min 1/ Standard <input type="checkbox"/> Min 5/ Cluster <input type="checkbox"/> Variety of Types		

**Tool Set #1 Appropriate Selection and Distribution of Assessments for an LAS
Template for Mathematics**

Grade Span (check one) <input type="checkbox"/> PK-4 elementary <input type="checkbox"/> 5-8 middle level <input type="checkbox"/> 9-12 secondary				Numbers and Operations Cluster			Shape and Size Cluster		Mathematical Decision Making Cluster			Patterns Cluster		
Assessment Title	When Given	Source of Assessment	Assessment Type	A	B	I	E	F	C	D	J	G	H	K
Total # of Assessments Minimum 8-12			Total # of Assessment Types Variety of types/cluster	# Measures ▪ Min 1/ Standard ▪ Min 5/ Cluster ▪ Variety of Types			# Measures ▪ Min 1/ Standard ▪ Min 5/ Cluster ▪ Variety of Types		# Measures ▪ Min 1/ Standard ▪ Min 5/ Cluster ▪ Variety of Types			# Measures ▪ Min 1/ Standard ▪ Min 5/ Cluster ▪ Variety of Types		

**Tool Set #1 Appropriate Selection and Distribution of Assessments for an LAS
Template for Science & Technology**

Grade Span(check one) ___ PK-4 elementary ___ 5-8 middle level ___ 9-12 secondary				Life Sciences Cluster			Physical Sciences Cluster			Earth and Space Sciences Cluster			Nature and Implications of Sciences Cluster					
Assessment Title	When Given	Source of Assessment	Assessment Type	A	B	C	E	H	I	D	F	G	J	K	L	M		
Total # of Assessments Minimum 8-12				Total # of Assessment Types Variety of types/cluster			# Measures ▪ Min 1/ Standard ▪ Min 5/ Cluster ▪ Variety of Types			# Measures ▪ Min 1/ Standard ▪ Min 5/ Cluster ▪ Variety of Types			# Measures ▪ Min 1/ Standard ▪ Min 5/ Cluster ▪ Variety of Types			# Measures ▪ Min 1/ Standard ▪ Min 5/ Cluster ▪ Variety of Types		

**Tool Set #1 Appropriate Selection and Distribution of Assessments for an LAS
Template for Social Studies**

Grade Span (check one) <input type="checkbox"/> PK-4 elementary <input type="checkbox"/> 5-8 middle level <input type="checkbox"/> 9-12 secondary				Civics and Government Cluster				History Cluster			Geography Cluster		Economics Cluster			
Assessment Title	When Given	Source of Assessment	Assessment Type	A	B	C	D	A	B	C	A	B	A	B	C	D
Total # of Assessments Minimum 8-12			Total # of Assessment Types Variety of types/cluster	# Measures ■ Min 1/ Standard ■ Min 5/ Cluster ■ Variety of Types				# Measures ■ Min 1/ Standard ■ Min 5/ Cluster ■ Variety of Types			# Measures ■ Min 1/ Standard ■ Min 5/ Cluster ■ Variety of Types		# Measures ■ Min 1/ Standard ■ Min 5/ Cluster ■ Variety of Types			

**Tool Set #1 Appropriate Selection and Distribution of Assessments for an LAS
Template for Career Preparation**

Grade Span (check one) ___ PK-4 elementary ___ 5-8 middle level ___ 9-12 secondary				Career and LIFE PLANNING		Career and LIFE MANAGEMENT	
Assessment Title	When Given	Source of Assessment	Assessment Type	Preparing for the Future A	Education/ Career Planning and Management B	Integrated and Applied Learning C	Balancing Responsibilities D
Total # of Assessments			Total # of Assessment Types	# Measures		# Measures	
Minimum 8-12			Variety per cluster	<ul style="list-style-type: none"> ▪ Min 1/ Standard ▪ Min 5/ Cluster ▪ Variety of Types 		<ul style="list-style-type: none"> ▪ Min 1/ Standard ▪ Min 5/ Cluster ▪ Variety of Types 	

**Tool Set #1 Appropriate Selection and Distribution of Assessments for an LAS
Template for Modern and Classical Languages**

Grade Span (check one) <input type="checkbox"/> PK-4 elementary <input type="checkbox"/> 5-8 middle level <input type="checkbox"/> 9-12 secondary				COMMUNICATION CLUSTER				CULTURE CLUSTER	
Assessment Title	When Given	Source of Assessment	Assessment Type						
Total # of Assessments Minimum 8-12				Total # of Assessment Types Variety per cluster		# Measures <ul style="list-style-type: none"> ▪ Min 1/ Standard ▪ Min 5/ Cluster ▪ Variety of Types 		# Measures <ul style="list-style-type: none"> ▪ Min 1/ Standard ▪ Min 5/ Cluster ▪ Variety of Types 	

**Tool Set #1 Appropriate Selection and Distribution of Assessments for an LAS
Template for Visual and Performing Arts**

Grade Span (check one) <input type="checkbox"/> PK-4 elementary <input type="checkbox"/> 5-8 middle level <input type="checkbox"/> 9-12 secondary				VISUAL AND PERFORMING ARTS CLUSTER		
Assessment Title	When Given	Source of Assessment	Assessment Type	Creative Expression A	Cultural Heritage B	Criticism and Aesthetics C
Total # of Assessments			Total # of Assessment Types	# Measures		
Minimum 8-12			Variety per cluster	<ul style="list-style-type: none"> ▪ Min 1/ Standard ▪ Min 5/ Cluster ▪ Variety of Types 		

Tool Set #1 Appropriate Selection and Distribution of Assessments for an LAS
SAMPLE* Template for Science & Technology *SAMPLE

Grade Span(check one)				Life Sciences Cluster			Physical Sciences Cluster			Earth and Space Sciences Cluster			Nature and Implications of Sciences Cluster			
<input checked="" type="checkbox"/> PK-4 elementary <input type="checkbox"/> 5-8 middle level <input type="checkbox"/> 9-12 secondary																
assessment Title	When Given	Source of Assessment	Assessment Type	A	B	C	E	H	I	D	F	G	J	K	L	M
Life Cycle Book (PK-2)	Spring, 1 st grade	MAP	Structured response	A3						D3				K6	L6	
Melts in the Sun (PK-2)	Spring, 2 nd grade	LAD	Scientific investigation					H1					J3	K3		
Insects and Me (PK-2)	after Insects Kit 2 nd grade	LAD	Research project			✓								✓	✓	
Energy Everywhere (3-4)	Winter 3 rd grade	LAD	Bundle					H1, H2								
Food for All (3-4)	Fall 3 rd grade	LAD	Bundle		B1, B2											
Science Around Us (3-4)	Winter 4 th grade	locally developed or adapted	Scientific critique				✓				✓	✓			✓	✓
Soils (3-4)	Spring, 4 th grade	LAD MMSA	Exhibition assessment								✓		✓	✓	✓	
Moving Massive Things (3-4)	Spring, 3 rd grade	locally developed or adapted	Scientific investigation						✓				✓		✓	
Plot Study (3-4)	Fall, 4 th grade	MAP	Scientific investigation	A1									K4	J1		
Earth's Movement (3-4)	Winter, 3 rd grade	LAD	Bundle									G3				
Total # of Assessments			Total # of Assessment Types 4 types/cluster 1 2 types/cluster 2 4 types/cluster 3 5 types/cluster 4	# measures 5 Min 1/ Standard Min 5/ Cluster Variety of Types			# measures 5 Min 1/ Standard Min 5/ Cluster Variety of Types			# measures 5 Min 1/ Standard Min 5/ Cluster Variety of Types			# measures 15 Min 1/ Standard Min 5/ Cluster Variety of Types			
10 assessments Minimum 8-12																

TOOL SET #2—Checking the Quality of Assessments

In Part I, we described the ways in which the quality of the assessments that a system comprises relates to the system’s reliability. There are two aspects of quality to be incorporated into the assessments selected for a system. These include clarity of language and clarity of scoring criteria.

Clarity of Language

This tool provides prompts for reviewing assessments with an eye on their language and structure and the extent to which they make expectations clear to students. Where the language or structure of an assessment is unclear or confusing, reliability is compromised.

Clarity of Scoring Criteria

Developing well articulated scoring criteria that minimize subjectivity in the scoring process contributes directly to reliability. The guidelines and templates in this tool set provide explicit steps for drafting and reviewing scoring guides. The “Guidelines for Drafting Scoring Guides” (page 45) provides an overview of relevant considerations for those developing scoring guides. The “Anatomy of a Scoring Guide” (page 44) identifies the important parts of a scoring guide, including the rubric, criteria, performance levels, sources of evidence, and performance descriptors. “Steps for Drafting Scoring Guides” (page 46) provides directions for completing a scoring guide template. We include two sample scoring guide templates as well as a sample scoring guide and scorer notes.

Tool Set #2 Checking the Quality of Assessments
CLARITY OF LANGUAGE

Review each assessment, paying particular attention to wording and layout, to ensure its clarity. The following prompts and suggestions for refinements are provided to assist in this process.

Are all expectations, including questions and directions,
clear and straightforward?

YES



Review the assessments for language and syntax.

NO



Replace vague or ambiguous expectations with more precise, explicit directions.

Are the language, syntax, and visual images familiar and
developmentally appropriate?

YES



Review the assessment's organization and layout.

NO



Replace unfamiliar terms, expressions, or images with language, syntax, or graphics more understandable and familiar to students.

Is the organization and layout straightforward and easy for
students to follow?

YES



The assessment's clarity will contribute to its reliability.

NO



Revise and/or reorganize the assessment so that expectations are clearly communicated and easy to follow.

Tool Set #2 Checking the Quality of Assessments
CLARITY OF SCORING CRITERIA

Anatomy of a Scoring Guide

	Performance Levels ↓			
Criteria ↓	1 ATTEMPTED DEMONSTRATION Does Not Meet The Standard	2 PARTIAL DEMONSTRATION Partially Meets the Standard	3 PROFICIENT DEMONSTRATION Meets the Standard	4 SOPHISTICATED DEMONSTRATION Exceeds the Standard
Content Standard: Performance Indicator: Source(s) of Evidence:	Performance Descriptor of a “1”	Performance Descriptor of a “2”	Performance Descriptor of a “3”	Performance Descriptor of a “4”

Note: NS (not scorable) may be assigned if the student response is:
 1. blank, 2. illegible, or 3. not responsive to the assessment.

- The **criteria** (content standards and specific performance indicators) and **performance levels** comprise the **rubric**.
- Ideally, this format and language remain consistent across the collection of assessments in a system.
- The **source of evidence** indicates the specific place(s) in the student work where evidence relevant to the criteria can be found.
- The **performance descriptors** provide detailed, task-specific descriptions of the kind, quantity, and quality of evidence necessary at each level of performance.

[The “note” in the table above indicates the option of assigning **Not Scorable** on any criterion.]

Tool Set #2 Checking the Quality of Assessments
CLARITY OF SCORING CRITERIA

Guidelines for Drafting Scoring Guides

- Criteria should be based directly on a performance indicator from Maine’s *Learning Results*.
- Descriptors should draw directly on the language of the content standard and performance indicator.
- Descriptors should refer to particular aspects of the assessment and, when possible, the rubric should list the “source of evidence” indicating where in the student work relevant evidence can be found.
- Descriptors should focus on the most important aspects of the performance indicator (content, concepts, and skills), not on the most easily measured.
- As much as possible, differentiation from one performance level to another should be made without reliance on subjective terms (e.g., “excellent”, “good”).
- Where necessary, include definitions of terms (this might include specific definitions for subjective terms) and/or scorer notes (guidance regarding correct/appropriate responses and/or background content information).
- Be absolutely clear about the use of **AND** and **OR** in descriptors, (e.g., “provides accurate definition for each term **AND** uses the terms appropriately in explanation of the solution” or “provides accurate definition for each term **OR** uses the terms appropriately in explanation of the solution”) .
- Revisit and refine scoring guides using student work to validate the descriptors.
- Include “not scorable” as an option in cases of blank, illegible, or off-topic responses.

Tool Set #2 Checking the Quality of Assessments
CLARITY OF SCORING CRITERIA

Steps for Drafting Scoring Guides

1. Identify/record the content standards and performance indicators under “criteria” in the rubric.
2. Identify the potential source(s) of evidence for each criterion (performance indicator) and record in the rubric.
3. For the first criterion (performance indicator), draft a descriptor for a “3” - meeting the standard.
4. Draft a descriptor for a “2” - partially meeting the standard. Consider all the ways a student might show partial demonstration.
5. Draft a descriptor for a “4” - exceeding the standard. To define exceeding or sophisticated demonstration, *consider*
 - a. levels of cognitive demand that exceed the expectation of the performance indicator,
 - b. performance indicators at the next grade span, and/or
 - c. the inclusion of relevant information and ideas that come from beyond the classroom experiences and indicate a sophisticated level of understanding or ability to apply.

Note: Several options are provided so that the one that is most appropriate for the performance indicator and the assessment can be selected. Review assessment to ensure that it provides an opportunity to demonstrate this level of performance.

6. Draft a descriptor for a “1” - does not meet the standard.
7. Repeat the process, as necessary, for each criterion (performance indicator).
8. Review the completed rubric and scoring guide against the Guidelines.

Scoring Guide Template – For Assessments with Multiple Performance Indicators

Assessment Title _____ Grade Level _____
Content Area _____

CRITERIA Content Standard (CS) and Individual Performance Indicators (PI) ↓	1 attempted demonstration (does not meet the standard)	2 partial demonstration (partially meets the standard)	3 proficient demonstration (meets the standard)	4 sophisticated demonstration (exceeds the standard)
CS: PI: Source of Evidence:				
CS: PI: Source of Evidence:				
CS: PI: Source of Evidence:				

Scoring Guide Template – For Assessments with a Single Performance Indicator

Assessment Title _____ Grade ____ Content Area _____

Content Standard and
Performance Indicator →

Source of Evidence:

4

sophisticated
demonstration
(exceeds the standard)

3

proficient demonstration
(meets the standard)

2

partial demonstration
(partially meets the standard)

1

attempted demonstration
(does not meet the standard)

SAMPLE "PATTERNS, RELATIONS, FUNCTIONS" Scoring Guide

K-2 Mathematics

	1 attempted demonstration (does not meet standards)	2 partial demonstration (partially meets standards)	3 proficient demonstration (meets standards)	4 sophisticated demonstration (exceeds standards)
G. Patterns, Relations, Functions 1. Recognize, describe, [extend, copy], and create a wide variety of patterns. Source of Evidence: Pattern Strip & Exceeds the Standard	The student attempts to create a pattern with objects.	The student creates at least a <ul style="list-style-type: none"> 3-element pattern and accurately represents it 2-3 times OR <ul style="list-style-type: none"> 2-element pattern and accurately represents it 3-4 times. 	The student <ul style="list-style-type: none"> creates at least a 3-element pattern (either repeating or growing) AND accurately represents it at least 4 times (there may be errors beyond the 4 times). 	The student <ul style="list-style-type: none"> creates at least a 3-element pattern AND accurately represents it at least 4 times (without any errors in entire sequence) AND accurately represents a 4-stage <u>growing pattern</u>.
G. Patterns, Relations, Functions 3. Represent and describe [both] geometric [and numeric] relationships. Source of Evidence: Questions 2, 6, 7, & Exceeds the Standard	The student attempts to <ul style="list-style-type: none"> use letters to represent the pattern or sequence OR <ul style="list-style-type: none"> describe how his/her creation is a pattern. 	The student accurately <ul style="list-style-type: none"> uses letters to represent the given sequence or sequence he/she created OR <ul style="list-style-type: none"> describes how his/her creation is a pattern. 	The student accurately <ul style="list-style-type: none"> uses letters to represent the given pattern AND the sequence he/she created AND describes how his/her creation is a pattern. 	The student accurately <ul style="list-style-type: none"> uses letters to represent the given pattern AND the sequence he/she created AND describes how his/her creation is a pattern, AND represents a growing pattern arithmetically.

Not Scorable should be assigned when student work is blank, illegible, or off task.

SAMPLE
“Patterns, Relations, Functions”
Scorer’s Notes

This sample is provided to illustrate the kinds of information (usually specific details about what constitutes a correct, appropriate, proficient response) that should be provided as scorer notes, in addition to the scoring guide.

For G1:

- The pattern needs to stop at the end of the pattern.
- The letter pattern needs to be complete (represents the 3 element pattern, 4 times).
- Growing patterns for example would show an increase or a decrease within the pattern.
- A growing pattern is not repeating and it can be increasing or decreasing (one or more elements within the pattern).

For G3:

- A student can use a two-element pattern for G3.
- A student’s letter pattern must accurately reflect the given patterns, not necessarily the entire sequence.
- The descriptions “it repeats” and “it keeps on” are acceptable.
- To Exceed the Standard: arithmetical representation will include operational symbols.

TOOL SET #3—Developing and Ensuring Scorer Accuracy

As discussed in Part I, interrater agreement plays a pivotal role in determining the reliability of the information provided by a local assessment system.

Beyond establishing the clarity and quality of scoring guides as addressed in Tool Set #2, there are several fundamental strategies for developing and ensuring reliable, valid scoring, or scorer accuracy. These include scorer training, scorer calibration, and scorer monitoring.

Scorer Training

Scorer training begins by communicating ground rules, expectations, and procedures for scoring student work. This sets the stage for achieving consistency in scoring. The training session also clarifies the specific expectations of the assessment to be scored and the details of the scoring guide to be used. Finally, the training includes the review of selected benchmarks/anchor papers, which are samples of student work that illustrate and illuminate the performance levels of the scoring guide and can inform scoring decisions.

Sample ground rules, bias considerations, “not scorable” rules, and guidelines for selecting and writing commentary for benchmarks are included in this tool set on pages 54 – 64.

Scorer Calibration

Before any scoring session, scorers must participate in calibration activities designed to standardize the application and interpretation of the scoring guide. This aspect of scorer training is essential to the achievement of scorer accuracy. Calibration includes individual scoring of common samples of student work. Subsequent discussions serve to verify points of agreement, clarify points of confusion, and resolve points of disagreement. Active participation in calibration activities is a key to the internalization of the standards of a scoring guide and consistency, or reliability, in scoring.

We recommend that *all* scoring sessions begin with calibration activities. This includes calibrating each day if scoring is completed on more than one day, and even each session, if scoring takes place both during a morning and an afternoon.

Guidelines for calibration sessions are included in this tool set on page 66.

Monitoring and Documenting Reliability

Interrater agreement, the most prominent estimate of reliability in a local assessment system, is monitored and documented through a system of double scoring. To achieve both purposes (monitoring for consistency and accuracy and documenting for verification), the following steps must be taken:

- Selection of a subset of student work to be independently scored by two different scorers

Identify a random, stratified sample of student work. See the table on page 70 for the appropriate sample size. Ensure representation of each teacher or classroom (stratification), but selected randomly within class sets.

- “score behinds” to monitor and provide opportunities to address levels of agreement during scoring sessions

Teacher leaders, or other experienced scorers who have a facilitation role, provide second scores for selected papers scored by each participating teacher. Where the two don't agree, the leader/facilitator provides feedback and clarification to the teacher. The purpose of this second scoring is *not* documentation of interrater agreement. The purpose is *monitoring* scorer accuracy and providing ongoing training and clarification as necessary.

- recording and comparison of the two ratings for each performance indicator

Using the template on page 74, the two scores from each double scored sample are recorded. This table will reveal, at a glance, levels of exact agreement. Likewise, the table will indicate particular criteria (performance indicators) or score points (1,2,3,or 4) where agreement is lacking.

- calculation of interrater agreement

Divide the number of scores with exact agreement by the total number of scores (see table of sample sizes on page 70) to calculate interrater agreement.

- follow up to address problems when interrater agreement falls below acceptable levels.

If interrater agreement does not meet acceptable levels (typically 70%), draw another sample for double scoring and repeat the procedure. If the double sample, in total, does not meet acceptable levels of agreement, then additional training and/or the refinement of the scoring guide are necessary and all of the papers must be rescored.

Tool Set #3 contains recommendations and templates to guide these aspects of scoring sessions.

**Tool Set #3 Developing and Ensuring Scorer Accuracy
Scorer Training**

Ground Rules for Scoring Student Work

Developing Shared Standards

Committing to a set of common guidelines helps to ensure consistency, and therefore reliability, in scoring student work.

- Believe that shared standards are possible.
- Agree to agree – work toward shared standards.
- Leave your personal standards at the door.
- Discussions are for clarifying, and are not arguments to win.
- Match evidence in student work to descriptor in scoring guide.
- Treat each criterion (performance indicator) separately.

Annotated Ground Rules for Scoring Student Work - Developing Shared Standards

Committing to a set of common guidelines helps to ensure consistency, and therefore reliability, in scoring student work.

- Believe that shared standards are possible.

Everyone involved in the scoring session must believe that consistent scoring is possible – regardless of the differences among the people involved in terms of background, philosophy, or style. Those who find this hard to believe, we ask to suspend their disbelief during the training and scoring sessions.

- Agree to agree – work toward shared standards.

Everyone has to work towards consistency. Playing the “devil’s advocate” will not help in this effort. We have to agree to make a good faith effort to interpret and apply the scoring criteria consistently.

- Leave your personal standards at the door.

Although everyone has their own personal and professional standards – expectations about what is “good enough” or appropriate, during the scoring session we must not apply our own standards, but rather the standards described in the scoring guide.

Personal standards are not “bad” – but they are not the standards we use when we are working for consistency and shared, common standards.

- Discussions are for clarifying, not arguments to win.

Particularly during calibration activities, people may disagree on particular scores.

Discussions that defend scores and point to the evidence that supports them are very helpful in understanding and internalizing the scoring guide. While these discussions may take on the tone of an argument, they are not won or lost. They serve a useful purpose in developing shared standards.

- Match evidence in student work to descriptor in scoring guide.

Scoring decisions must be made based on particular evidence in the student work and its correspondence to particular descriptors in the scoring guide. This is what we all have in front of us and what will allow us to score consistently. It is a good habit to remember that you should be able to point to the evidence in the student work and to the language in the descriptor on the scoring guide. If you find yourself referring to what you “know” or “think” the student meant, then you are not matching evidence to the descriptor.

- Treat each criterion (performance indicator) separately.

We need to find different evidence for each score given, and not let a student’s performance on one criterion influence our judgment on a different criterion. This is a type of bias, which we will talk more about, and which we must avoid.

**Tool Set #3 Developing and Ensuring Scorer Accuracy
Scorer Training**

Scorer Bias Considerations

Scorer bias refers to any feature in student work or any aspect of a scoring situation, which may influence scoring decisions. Scorer bias may lead to artificially inflating or depressing scores. Scorers must guard against bias and assign scores based only on the evidence found in the student work and the criteria and descriptors in the scoring guide.

Potential Sources of Scorer Bias

- Appearance of work - neatness or messiness/legibility
- Personal reaction to topic, subject, strategy, reference
- Tone of student communication
- “Halo” effect - strong performance in one criterion influencing the scoring of another aspect of the work, or the reverse - poor performance in one criterion influencing the scoring of another aspect of the work
- Apparent effort or improvement from previous efforts
- Length or complexity of student response
- Relative quality (i.e., “better than others I’ve seen”)
- Familiarity with the student

Annotated Scorer Bias Considerations

Scorer bias refers to any feature in student work or any aspect of a scoring situation, which may influence scoring decisions. Scorer bias may lead to artificially inflating or depressing scores. Scorers must guard against bias and assign scores based only on the evidence found in the student work and the criteria and descriptors in the scoring guide.

Potential Sources of Scorer Bias

- Appearance of work - neatness or messiness/legibility

It is common to equate neat work with good quality and to assume that messy (hard to read or hard to follow) work is of poor quality. This isn't necessarily true. Unless the scoring guide includes specific expectations for the appearance of the work, it should not influence scoring decisions.

- Personal reaction to topic, subject, strategy, reference

A student's work may evoke a personal reaction – if he or she has written a paper supporting a position that you don't share, or if he or she has selected a subject that you don't find interesting or important. Your reaction should not influence your scores. Likewise, a student may use a strategy that you find inefficient, or an approach that you don't recommend. Unless the scoring guide specifies a particular strategy or approach, this should not influence your scores. Finally, a student may refer to a resource, incident, person, or idea that you have a personal reaction to. Again, this should not influence your scores.

- Tone of student communication

Sometimes student work has “voice” and we can sense how the student felt about the assessment or the subject. This should not influence scoring decisions. A student may be very positive and fail to provide evidence of proficiency. Likewise, a student may display a negative attitude but include all of the evidence necessary to meet or exceed standards. The student's tone or attitude should not influence scoring.

- “Halo” effect - strong performance in one criterion influencing the scoring of another aspect of the work, or the reverse, poor performance in one criterion influencing the scoring of another aspect of the work.

When a piece of work is scored on more than one criterion, it's important to treat each one separately. It is possible for a student to score very well on some aspects of an assessment and poorly on others.

- Apparent effort or improvement from previous efforts

Effort is not a part of what we are scoring. Generally you assess and provide feedback on effort in your own classrooms. When we score using a scoring guide, your sense that a student has worked very hard, should not influence your score, nor should a sense that a student should have tried harder.

- Length or complexity of student response

Length does not necessarily correspond to quality, nor does complexity. Don't assume that longer responses are "better" or that short responses are not good enough. Look at the actual evidence and make the scoring decision based on its correspondence to the scoring guide.

- Relative Quality (i.e., "better than others I've seen")

Standards-based scoring makes no provision for relative quality but it is easy to get "excited" about a piece of student work that seems better than others. Don't assume that it deserves any particular score (a 3 or 4). You still have to look at the evidence and the scoring guide and make the decision.

- Familiarity with the student

If you are scoring your own student, or students with whom you are familiar, you may have preconceived notions of what they can, or should, or will do in terms of the assessment that you are scoring. You must set these notions aside. You cannot assume that a student who usually does good work will do well on this assessment and you cannot assume that a student who often struggles will have difficulty with this assessment. As we've said before, you must score the work based on the evidence and the scoring guide.

**Tool Set #3 Developing and Ensuring Scorer Accuracy
Scorer Training**

“Not Scorable” Guidelines

In addition to the four levels of performance described in the scoring guide, there is an option to assign NS, or “not scorable”, on any criterion or on an entire piece of student work.

NS is assigned to student work *only* for one of the following three reasons:

- no evidence available – no response, blank
- student work is illegible – can’t be deciphered
- student work is completely off task –
not responsive to the prompt or problem

Note: NS is considered a “score” and interrater agreement must be established. When NS is assigned, the student work should be treated like any other piece in terms of the double scoring process.

Annotated “Not Scorable” Guidelines

In addition to the four levels of performance described in the rubric and scoring guide, there is an option to assign NS, or “not scorable” on any criterion or on an entire piece of student work.

NS is assigned to student work ***only*** for one of the following three reasons.

- **no evidence available – no response, blank**

If all of an assessment is blank, or all parts of an assessment necessary for a particular criterion (performance indicator), are blank, then there is no basis to make a judgment about the quality of performance. A scorer cannot determine whether or not the student attempted the assessment, was present when the assessment was administered, or saw the question and therefore cannot score the work.

- **student work is illegible – can’t be deciphered**

If the entire response to an assessment is illegible, or all parts of an assessment necessary for a particular criterion (performance indicator) are illegible, such that the work cannot be deciphered, then there is no basis to make a judgment about the quality of performance. A scorer cannot determine how well the evidence matches descriptors in the scoring guide and therefore cannot score the work.

- **student work is completely off task –not responsive to the prompt or problem**

If the response to an assessment is off task, or all parts of an assessment necessary for a particular criterion (performance indicator), are completely off task and do not to respond to the assessment, there is no basis to make a judgment about the quality of performance. A scorer cannot determine whether or not the student saw, read, or understood the assessment and therefore cannot score the work.

All three reasons for assigning NS are based on an understanding that when a judgment cannot be made, it is fairer and more accurate to call the work “not scorable” and to require the student retake the assessment, or another version of it. When a judgment can be made, when there is evidence that the student saw, read, and understood what was being asked for, then the appropriate score of 1-4 must be assigned.

**Tool Set #3 Developing and Ensuring Scorer Accuracy
Scorer Training**

QUICK TIPS FOR SCORING STUDENT WORK

- **Review the task and the student work.**
- **Focus on the first criterion, read the level 3 “proficient” descriptor.**
- **Look for evidence of that standard in the student work – note the recommended “source of evidence”.**
- **Assign a score as follows:**

Performance Level →	1 attempted demonstration little evidence	2 partial demonstration some evidence	3 proficient demonstration evidence meets standards	4 sophisticated demonstration evidence exceeds standards
Performance Indicator	if yes assign 1	if yes assign 2	START HERE	if yes, assign 4
Source of Evidence:	if blank, illegible, or not responsive to the assessment, assign NS	if no, check 1 ←	if yes, check 4 → if no check 2 ←	if no, assign 3

- **Record your score (*noting evidence to support it*).**
- **Repeat the process for remaining criteria.**

Tool Set #3 Developing and Ensuring Scorer Accuracy Scorer Training

GUIDELINES FOR SELECTING BENCHMARKS

Ideally, this work is done as part of an assessment’s field test. After double scoring all of the student work produced during the field test (as many as 75-100 pieces), establishing the potential reliability of the assessment, and finalizing the descriptors in the scoring guide, benchmarks can be selected.

Benchmarks and supporting commentary should be provided as part of scorer training.

- Sort all scored student work into two piles, one for “Potential Benchmarks” (all papers with exact agreement across criteria) and the other for “All Others”. *Papers with exact agreement are preferred as the evidence for the assigned score is usually more apparent.*
- Sort all of the papers in the potential benchmark pile into four piles (1s, 2s, 3s, and 4s).
- Look through the “1” pile and identify the sample(s) with the most 1s assigned *across criteria*. Then identify the paper within that set that has the highest interrater agreement *across criteria*.
- Repeat for 2s, 3s, and 4s.
- If more than one paper is identified for a performance level, discuss and decide on the benchmark sample based on reflection of scoring guide descriptors and/or clarity and legibility of the student work.. *Benchmarks should, first and foremost, exemplify the performance levels but they must also reproduce well in order to be useful.*
- If there are no papers with exact agreement at a specific score point, use information from third scoring.
- Goal – one paper at each score point with, as much as possible, the same score assigned across criteria. *The goal of one paper at each performance level across criteria does not indicate a desire or expectation for homogeneous performance across criteria by individual students. Rather, it is an efficiency in the preparation and use of benchmarks.*
- Draft commentary using the guidelines and template provided.

**Tool Set #3 Developing and Ensuring Scorer Accuracy
Scorer Training**

**Guidelines for Writing Commentary or
Scoring Rationales for Benchmarks or other Training Materials**

Benchmark samples of student work may be used for a variety of purposes during professional development experiences. These include:

- To illustrate the levels of performance on a scoring guide and to illuminate the distinctions among levels.
- To refer to during scoring activities, in combination with the scoring guide, in order to assign appropriate scores.
- To illustrate the attributes of scorable and unscorable tasks or assignments.

Additional samples of scored student work can be used

- To provide opportunity for additional scoring training, calibration, and/or “qualifying” as a reliable scorer.

In order to achieve any of these purposes, the pre-scored pieces should be accompanied by clear, detailed commentaries. The commentary explains the assigned score by providing a description of the specific evidence identified in the student response and the ways in which it corresponds to a described level of performance on the scoring guide/rubric.

A good commentary or scoring rationale is one that:

- is prepared with the purposes discussed above in mind
- refers to specific evidence in the student work
- refers to specific language in the scoring guide
- explains, in cases of “borderline” papers, why the assigned score is more appropriate than the score point above or below
- anticipates possible points of disagreement and provides evidence to refute them and
- is clear, thorough, and detailed.

Tool Set #3 Developing and Ensuring Scorer Accuracy
Scorer Training

Template for Scoring Rationale/Commentary

Criteria	Score	Rationale Commentary Evidence and Explanation

***SAMPLE* Scoring Rationale /Commentary**

Assessment Title: Patterns

Subject Area: Mathematics

Identification: BENCHMARK 3

Grade: PK-2

Date: 3/20/04

Criteria	Score	Rationale Commentary <i>Evidence and Explanation</i>
G. Patterns, Relations and Functions 1. Recognize, describe, ... and create a wide variety of patterns.	3	The student created a three element pattern by drawing “triangle-triangle-trapezoid, rhombus” four times.
G. Patterns, Relations and Functions 2. Represent and describe ... geometric relationships	3	The student accurately named the given pattern “ABC”. The student also accurately named the pattern he/she created “AABC” and explained why his/her creation is a pattern, “because it repeats (repeats) itself (itself).”

**Tool Set #3 Developing and Ensuring Scorer Accuracy
Scorer Training - Calibration**

Facilitating Training and Calibration Sessions

Initial scorer training usually includes a review of the scoring guide and of the benchmarks and the benchmark commentary. This should be followed by practice scoring and calibration. A successful calibration session will increase the consistency and accuracy of scoring. During calibration, the scoring guide may be refined to add details or terms that will support consistent judgments by scorers. These changes may not alter the expectations of the assessment; they merely clarify the performance levels. In addition, scorer notes that will inform scoring decisions may be developed or embellished.

A calibration session is the opportunity for scorers to develop a common interpretation of the standards. Independent scores are discussed with evidence in the student work supporting application of the scoring guide. Scorer notes are developed to assist in the common interpretation of the scoring guide and student evidence. The purpose of the calibration session is not to refine the task itself, but to clarify and illuminate the scoring process.

Tools and Materials for a session when prescored papers are not available:

- ✓ Chart paper and/or packet with copies of
 - Ground Rules,
 - Scorer Bias Considerations,
 - Using a Scoring Guide,
 - “Not Scorable” Guidelines,
 - Scoring Guide, and
 - Benchmarks with commentary.

- ✓ Multiple copies of student work.. Randomly select 5 (+/-) total pieces of student work (if the task has been administered in multiple classrooms, then select 1-2 from each teacher) and make photocopies of these for each scorer.
- ✓ Post-It notepads.
- ✓ Optional - Computer and printer access for scoring guide and/or scorer note refinements

Procedure:

1) An administrator, teacher leader, or other willing participant assumes the role of facilitator; another assumes the role of recorder (to document noted changes for scoring guide and scorer notes).

2) The facilitator leads a review of ground rules, scorer bias considerations, “not scorable” guidelines and the scoring guide.

See the annotated versions of these documents. When reviewing the scoring guide, point out the performance indicator(s) being scored and the source of evidence for each. Review the descriptors for the performance levels.

3) The facilitator leads a review of the benchmarks and the benchmark commentary.

Discuss the specific evidence found at each performance level on each criterion.

4) The facilitator distributes a copy of student work sample #1 to each scorer.

5) Each scorer independently reviews the student work, assigns a score for each performance indicator, and notes evidence supporting his/her score(s).

6) The facilitator leads a discussion.

- How many scored this piece a “3?”; “4?”; “2?”; “1?”

Ask for a show of hands for each score.

- Would anyone like to share the evidence in the student work that led to the score he/she assigned?

*Make sure that scorers refer to evidence in student work, **not opinions**. Ask for clarification when necessary. Make sure to refocus the group to the evidence or score being discussed when necessary.*

The discussions supporting scoring decisions should focus on evidence in the student work, language in the scoring guide, and specific details in the benchmarks and the benchmark commentary.

- As a group, agree on the correct interpretation of the evidence according to the scoring guide.
- The recorder can document a scorer note(s) to clarify decisions reached based on the discussion.

The facilitator continues this process for another 3 or 4 pieces of student work until independent scores are in agreement.

At the end of the calibration process, a printed copy of any additional scorer notes is given to each scorer along with a refined scoring guide (if produced).

7) The group scores one more common piece of student work using the revised scoring guide and scorer notes.

8) If independent scores are not in agreement continue this calibration process by scoring common samples until consistency is achieved.

9) The facilitator reviews procedures for double scoring and the importance of independent scores for interrater reliability before the scoring session begins.

10) If more than one scoring session is necessary (morning and afternoon or on two or more days), one or more additional calibration exercise should be completed at the beginning of each session.

If collections of prescored student work are available, calibration activities proceed in a manner similar to the one described above, but rely on the scores and commentary provided in the prepared calibration materials.

**Tool Set #3 Developing and Ensuring Scorer Accuracy
Monitoring and Documenting Interrater Agreement**

Guidelines for Selecting Student Work for Double Scoring

- Using the table provided on the next page, identify the appropriate sample of student work. The targeted “adequate” sample sizes in this table provide confidence that the calculated interrater agreement is indicative of the degree of scoring consistency among all papers (rather than merely those that are double scored).
- Ensure that the sample of papers to double score is representative, drawing from each class or teacher. If the sample of double scored papers is not representative, the accuracy of the resulting agreement data is called into question.
- Select individual pieces randomly. Divide the total number in the desired sample by the number of classes or teachers to be represented. Randomly select that number of pieces from among all of the pieces produced in a particular class or classroom.
- Distribute papers to be double scored among all papers to be scored. Establish a system that will ensure that all participating scorers are represented in double scoring and that double scoring takes place throughout the scoring session.

Note: Double scoring to calculate interrater agreement provides after-the-fact information about reliability of scoring. In order to obtain ongoing information about consistency among scorers, one or more “expert,” experienced scorers should “score behind” all members of the scoring group. This provides consistency checks and allows for necessary clarifications and retraining for individuals, as necessary. When a “score behind” reveals disagreement, that should be noted for overall interrater agreement data, but an effort must be made to determine the piece’s “true” score to assign to the student.

Strategic use of “score behinds” and “spot training” can greatly enhance scorer consistency and accuracy.

**Tool Set #3 Developing and Ensuring Scorer Accuracy
Monitoring and Documenting Interrater Agreement**

Appropriate Sample Size & Interrater Agreement for Double Scoring

Total Number of Papers	Adequate Sample for Double Scoring	Exact Agreement of 70% or better	Exact Agreement of 70% or better across 2 samples (if necessary)
1-25	10	7/10	14/20
26-74	17	12/17	24/34
75-250	25	18/25	35/50
over 250	10% of all papers	70% of double scores	70% of double scores

Depending on the total number of students completing the assessment, identify the appropriate sample for double scoring.

After double scoring the sample, check interrater agreement.

Adequate interrater agreement in the double scored sample verifies the reliability of the scores produced by the assessment.

If interrater agreement is *not* adequate, identify a second sample for double scoring.

Adequate interrater agreement across the two samples of double scored papers verifies the reliability of the scores produced by the assessment.

If interrater agreement still is *not* adequate, see recommendation on page 75.

**Tool Set #3 Developing and Ensuring Scorer Accuracy
Monitoring and Documenting Interrater Agreement**

Double Scoring Student Work

It is important to assign scores to student work *independently*.

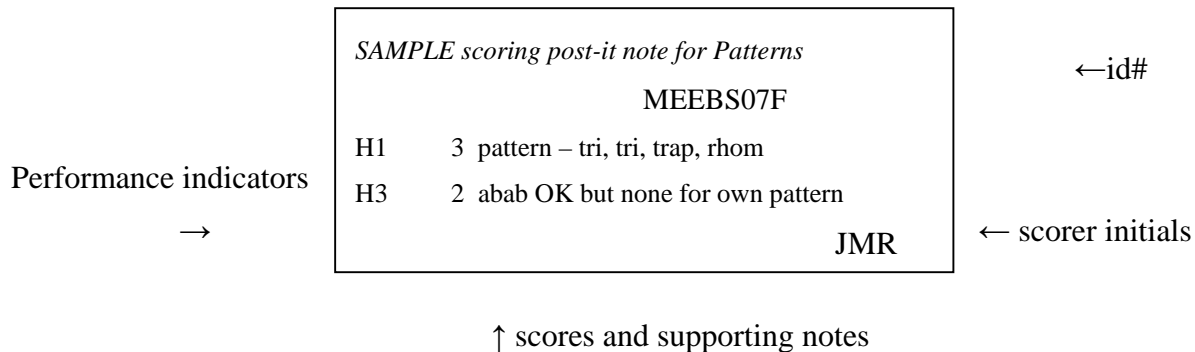
During double scoring, two people will score a piece of student work. Each scorer will assign a score without either discussing the work with others or being influenced by the score assigned by another.

If you are the first scorer,

- Read the work and refer to the scoring guide and benchmarks
- Assign the necessary scores on a post-it note.

It is helpful if you write the letter and number of the performance indicators (e.g., F1, L2) and then the assigned score beside it (F1 - 3, L2 – 2). If the piece is scored for only one performance indicator, writing the letter of the performance indicator is not necessary.

- Write your initials on the post-it note with your scores and remarks supporting scores.
- Cover the post-it note with a blank note and
- Write your initials on this top post-it note *this will keep you from selecting this work to score as a 2nd scorer.*
- Place the student work in a pile to receive a second score.



If you are a second scorer,

- Read the work and refer to the scoring guide and benchmarks
- Assign the necessary scores on the post-it note placed on top of the first scorers' post-it and scores (don't peek!).

It is helpful if you write the letter of the performance indicators, and then the assigned score beside it, F1 - 3, L2 - 4.

- Write your initials on the post-it note with your scores.
- Place the student work in a pile for work that has been scored twice

Guidelines for Establishing Reliability for Assessments involving Observation, Presentation, or Performance

As with all assessments in the LAS, when scoring assessments based on performance – an observation checklist, exhibition, or other presentation, reliability can be established through interrater agreement or documentation of individual scorer calibration.

Interrater agreement can be established in one of the following ways:

- Two scorers make the observations or listen to the presentations of the appropriate sample of students, score independently and calculate their agreement. The second scorer may be another teacher of the same grade level or discipline (2 third grade teachers scoring students on an observation checklist for science procedures), a teacher from another grade span (a high school physical education teacher spending a day scoring with an elementary physical education teacher), or another educator (an assistant principal and an eighth grade English teacher listening to and scoring oral presentations by a middle school English class).
- Videotape the sample of assessments to be double scored for a second scorer to review and score at a later date. Record first and second scores and calculate agreement. (every teacher videotapes two social studies presentations and brings them to a scoring session where teachers exchange videotapes and provide second scores to one another's students).

The other option for documenting reliability is becoming “certified” as a scorer. The teacher administering and scoring the assessment completes training activities and scores a calibration set which has already been assigned “true” scores. If the teacher demonstrates 80% or better agreement with the “true” scores, he or she has documented individual scorer calibration and can score all of his or her own students.

Tool Set #3 Developing and Ensuring Scorer Accuracy
Monitoring and Documenting Interrater Agreement

***SAMPLE* Interrater Agreement Tally**

Grade Span: K-2

Discipline: Mathematics

Assessment: Patterns

Criterion: G1

Scores Assigned by Rater 2	Scores Assigned by Rater 1					
		NS	1	2	3	4
	4					II
	3			III	IIIIII	I
	2		I	III	II	
	1		III			
	NS	I				

To calculate interrater agreement, divide the number of exact agreements (tallies in shaded boxes) by the number of pieces double scored and multiply by 100.

In the example above, there are eighteen papers with exact agreement from twenty five papers double scored, $18/25 = .72 \times 100 = 72\%$.

Tally and calculate interrater agreement for each criterion.

Note: a copy of this tool will be necessary for each criterion on each assessment. Multiple grids can be reproduced on a single piece of paper, two to a side.

Interrater Agreement Tally

Grade Span:

Discipline:

Assessment:

Criterion:

	Scores Assigned by Rater 1					
Scores Assigned by Rater 2		NS	1	2	3	4
	4					
	3					
	2					
	1					
	NS					

Interrater Agreement Tally

Criterion:

	Scores Assigned by Rater 1					
Scores Assigned by Rater 2		NS	1	2	3	4
	4					
	3					
	2					
	1					
	NS					

**Tool Set #3 Developing and Ensuring Scorer Accuracy
Monitoring and Documenting Interrater Agreement**

Recommendations for Addressing Insufficient Interrater Agreement

Review levels of interrater agreement for each criterion scored.

- If interrater agreement is 70% or better, it is considered reliable for the purposes associated with local assessment systems: informing teaching and learning, monitoring programs, and certifying achievement for graduation.
- If interrater agreement is between 65% and 70%, it may be reliable enough within the context of the LAS. If most (more than three quarters) of the ratings included in the system have demonstrated reliability of 70% or above, levels of agreement between 65% and 70% may be considered satisfactory and will not reduce the overall consistency of the LAS below acceptable levels. Caution should be used, however, in drawing conclusions based on these ratings as a single measure.
- When interrater agreement falls below acceptable levels, even after a second sample had been double scored (see page 70), rescoring will be necessary. This must be preceded by steps designed to improve the levels of agreement in the next set of ratings. These steps might include:
 - Examining the Interrater Agreement Tallies to look for issues – particular score points where agreement is low and clarify the scoring guide accordingly.
 - Providing additional scorer training, focusing on identified areas of disagreement and providing additional samples to illustrate and illuminate points of differentiation on the scoring guide.
 - Revisit and re-stress ground rules and sources of bias with individuals or all scorers, as necessary.
- All papers must be rescored on each criterion with inadequate agreement. A different sample for double scoring must be identified and double scored and interrater agreement must be calculated and reviewed as described above. Only the ratings from the second (or reliable) scoring session may be used to report on student performance.